Supporting Teachers in Higher Education in Designing Test Items

Draaijer, S.

2016

document version
Publisher's PDF, also known as Version of record

Chapter 8: Summary and Discussion

8 Summary and Discussion

The work accomplished in this thesis responds to the two main aims concerning the design process of test items for in-house developed quizzes and exams by teachers in higher education. The initial aim was to develop interventions in the form of methods and techniques for teachers, to support them in improving the test item design process. To establish the extent of their appropriateness and effectiveness, scientific studies were conducted. This need for the development of support stems from an engineering focus on solving real-world problems. The initial aim presented in this thesis was necessarily preceded with the aim of which to understand the test item development task better and develop a model of the cognitive processes involved in test item design on which the developed support should be based.

In this final chapter, an overview of the key findings is presented, followed by an integration of findings according to the main themes of the dissertation. Limitations and recommendations for further research are formulated.

8.1 Summary of research questions and findings

Developing a process model of test item design and generating interventions for test item generation was driven by answering a set of research questions. Each research question was addressed in a separate paper presented in the Chapters 2 to 7 of which the summarized findings will be presented next.

8.1.1 Research question 1 (Chapter 2)

> *What are the characteristics of a process model for test item design that is based on the concept that the test item design task is an ill-defined creative design problem-solving task?*

In Chapter 2, a critical appraisal was put forward regarding the limited usefulness of current literature for test item writing for teachers in higher education and the need for a different perspective on the test item design task. Therefore, the position was taken to consider test item design as a form of problem solving. It was

concluded that designing test items fits well within the concept of solving ill-defined design problems. For such problems, a descriptive model based on the creative problem-solving approach (Isaksen & Treffinger, 2004) is suitable; in Chapter 2, this model was particularized for the test item design process from a cognitive perspective. The process model developed makes divergent and convergent phases central to the development process. See Figure 1 for a representation of the process as developed.

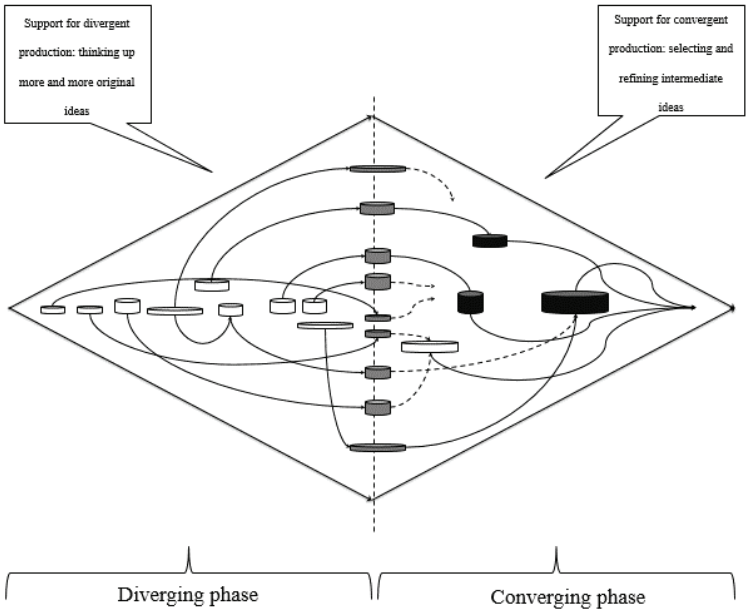

Figure 1. Representation of diverging and converging phases in test item development. Cylinders represent test items at different stages and with various technical and originality measures.

In the model, divergent production was visualized by outward-bound arrows that indicated a growth in both the number and originality of developed test item ideas. The model reflected how the attention of the designer moves back and forth from problem to solution, how solutions come into existence and evolve, what the breadth and width of the solution space is, and a general sense of the number of

solutions that could be developed. The model was able to describe the behavior of both novices and experts.

The model could also be used prescriptively among both novice and expert teachers by *stimulating* the teacher into intentionally divergent and convergent thinking and production of test items. This approach could lead to a higher yield in terms of number of test item ideas and higher quality in terms of more original or more technically sound ideas. The model showed that interventions to stimulate the number and quality of test items envisioned could be geared towards improving the diverging and converging process during test item design and hence provide a route for the development of such interventions.

8.1.2 Research question 2 (Chapter 3)

Which practical inspirational guidelines to support divergent production of test items for teacher in higher education can be identified and what is their effectiveness.

In Chapter 3, practical compact inspirational guidelines that teacher could use to support them to spark imagination for ideas for test items were identified and described. Such compact guidelines could serve as support interventions for the developed process model described in Chapter 2. These inspirational guidelines were identified on the basis of experiences of participants in test item design projects and literature. The appropriateness and effectiveness of these guidelines was evaluated on the basis of several test item development projects a case studies. The study resulted in a list of 60 guidelines that were grouped in 10 categories: seven categories consisted of guidelines that tap into the use of experiences and available resources for test item designers and three categories were regarded to be essentially traditional requirements for test item design.

Based on test item development case studies it was concluded that in many situations in which teachers have to design test items, at least four guidelines were found useful for their particular context or situation. Further, teachers stated that specialized support was needed to scaffold them in using the guidelines.

8.1.3 Research question 3 (Chapter 4)

Which design patterns to support divergent production of selected response test items can be identified in teacher-generated test items, and how can they be described?

In Chapter 4, the concept of design patterns was studied. Design patterns are generic combinations of solutions to recurring problems within problem-solving or design domains. As such, design patterns could also form a support intervention for the developed cognitive process model as developed in answer to research question 1. Expert designers can, by means of that expertise, rapidly match a problem to the appropriate design pattern to arrive at satisfactory solutions to given problems and contexts. Design patterns are in that sense culminations of schematic cognitive data based, often idiosyncratically, on the particular expertise of an individual.

Templates for design patterns for test items were developed and prototypical test items to accompany the design patterns were identified. Specific design patterns were identified that aim to query for higher-order learning outcomes by matching particular content and arrangements of on-screen elements with test item stimuli and responses. About 30 design patterns emerged, of which 10 were fully described. A template for a design pattern contained the descriptive characteristics shown in Table 1.

Table 1.

Characteristics of design patterns.

Name	Description
Title	Overall description of the design pattern
Context	Indication of type of learning objective, domains of interest, etc.
Knowledge Skills and Abilities (KSA)	Indication of type of learning objective to be measured
Pattern Core	Description of the most important elements of the stimulus, responses, and on-screen elements
Design Effort	Indication of time needed to arrive at a main idea for a test item
Realization Effort	Indication of time needed to produce and test the item in detail
Extraneous Cognitive Load	Unwanted cognitive effort to understand the content of the test item
Guess Chance	Probability of getting an test item correct by mere chance
Iconic Examples	Quintessential instances of the design pattern
Scoring Rules	Rules for awarding defensible scores to the test item

8.1.4 Research question 4 (Chapter 5)

What are the main requirements for a computer support tool to support divergent and convergent production of selected response test items, and how can these requirements be met by a computer application?

Research question 4 dealt with the problem of how the ill-defined problem-solving character of test item design and the process of diverging and converging could best be supported by an online computer tool. As the guiding principle for that tool, the concept of a *virtual workbench* (Hewett, 2005) was chosen. A concept of the tool was developed and arguments were put forward regarding the appropriateness of the tool, its functions, incorporated resources and user

interaction design. The first main distinguishing characteristics of the tool with respect to similar efforts were the flexibility of the tool as promoted in the literature regarding computer tools to support creative problem solving. The second distinguishing function was that the system provided support information and prompts to engage in deliberate divergent and convergent production of test item content. The support information was aligned to the phase in which a teacher is active and the specific test item format. The strong points of the system are that it brings support information as near to the teacher as possible during the act of designing test items, yet remains unobtrusive.

8.1.5 Research question 5 (Chapter 6)

To what extent can divergent and convergent production of test items be improved by encouragement to diverge, developing a concept map, presenting item shells, and presenting classic test item writing guidelines?

In Chapter 6, an experimental study using a multilevel between-subjects design was presented, in which two specific support interventions to improve the divergent and convergent production of test items were studied. The intervention to stimulate divergent production and increase originality of test items consisted of explicitly encouraging participants to diverge, developing a concept map, and making item shells available via the online computer support tool. The intervention to stimulate the convergent production of test items and increase the technical quality of test items consisted of presenting often-used convergent guidelines from Haladyna, Downing, and Rodriguez (2002) for clear and unambiguous test items.

The analysis showed no significant effect of the intervention in stimulating divergent production of ideas with respect to the *number* of test items developed. However, analysis indicated that the intervention to stimulate divergent production did have an effect on the degree of originality of developed test items. No effects were observed of an increase in technical quality of those items. However, designing more test items resulted in lower levels of technical quality.

8.1.6 Research question 6 (Chapter 7)

To what extent does the presentation and use of an idea leaflet result in test items that are more original as compared to not using an idea leaflet,

a) when the idea leaflet is presented prior to a test item design task and used in the first phase of idea generation?

b) when the idea leaflet is presented and used after initial exhaustion of ideas?

Chapter 7 dealt with an experimental between-subjects study in which a more general creativity intervention, called an idea leaflet, to stimulate the production of original test items was studied. The study also took the variable of expertise and the concept of exhaustion of ideas into consideration.

First, the rationale for the idea leaflet and its content was developed. The main consideration in the rationale was that perspective shifts by the teacher are required in order to produce test items with a higher degree of originality. The idea leaflet was supposed to promote the occurrence of such perspective shifts an lead to more original test items.

The study showed that the intervention resulted in a *higher degree of originality* of developed test items, provided that the idea leaflet was presented after some development time in the phase of extended effort. The study was in line with findings from creativity studies showing that participants start creativity tasks by writing down the more obvious ideas; only after sustained effort do they generate more unusual ideas. Active renewed processing of the support was shown to be needed to lead to an increase in originality after initial exhaustion of ideas.

8.2 General conclusion

The most important question regarding any thesis is what the findings add to the knowledge regarding a particular problem or phenomenon. What is the intrinsic contribution of this thesis?

The main contribution of this thesis is that it establishes the value of regarding the task to develop test items as solving an ill-defined design problem in which divergent and convergent thinking and production play essential roles. Regarding the task as an ill-defined problem-solving task enables scholars to describe and understand this task more accurately and to develop better support for higher education teachers in increasing the quality of test items.

In particular, divergent thinking and production is acknowledged to be vital for producing new and original test items. Emphasizing divergent production and establishing deliberate interventions to support divergent production of test items results in more original test items. As the developed interventions are compact and easily accessible, it is likely that they will be used and lead to improved results in the practice of higher education.

8.3 Implications for theory

In terms of scientific significance, the study adds new insight into theory building and empirical findings to the existing literature regarding test item design and computer support tools for test item design.

First, the studies in this thesis connect the design of test items to general theories of design, problem solving, and expertise in an interdisciplinary fashion. For the first time, test item design is described elaborately from the cognitive perspective of the teacher. More light was shed on Haladyna's observation that "The cognition that is used in writing the item is still a mystery" (personal communication, May 15, 2013). Taking the teachers' cognitive perspective as a starting point and building a cognitive process model led to a new and richer appraisal of the test item design task, resulting in an enhanced description of the complexity of that task, a complexity of which only some findings had been delineated in the classic literature regarding test item design.

Second, from the perspective of theory building for test item design, the thesis adds a new complimentary model for the design process. With that new model, it is not the abstractly formulated goals of achievement tests that are central, but rather the domain knowledge and skill of the teachers combined with purposeful divergent

and convergent thinking, the instructional materials, and the instruction provided to the student. Figure 2 shows a visual representation of the position a teacher occupies theoretically within the concepts that surround designing test items.

Third, as the test item design task was positioned with literature regarding design and ill-defined problem solving, the thesis also extends this body of knowledge by studying a problem type that forms a gap in problem types that have been researched. This gap lies in the continuum of well-definedness versus ill-definedness on the one hand and small versus large and complex on the other. The nature of solving well-defined and smaller problems such as math problems, reasoning problems, figural problems, and physics problems (Chi, Feltovich, & Glaser, 1981; Davidson & Sternberg, 2003) has been studied extensively. Learning to solve large and complex physical and non-physical artifacts such as houses, war ships, products, computers, instruction, services, or political problems (Cross, 2007; Jonassen, 1997; Jonassen & Mandl, 1990; Rowland, 1993), or naturalistic decision making problems with incomplete information (Lipshitz, Klein, Orasanu, & Salas, 2001) have also received extensive attention. However, designing in an ill-defined context with relatively small, non-physical artifacts that have to be designed repeatedly from scratch yet must comply with many restrictive rules, has not been studied in depth. Comparable problems of this type are writing columns or articles for magazines (Jacobi, 1997), making three-panel comics for magazines, or writing blog messages. This thesis could add insight to learning to solve and solving those creative problems types as well.

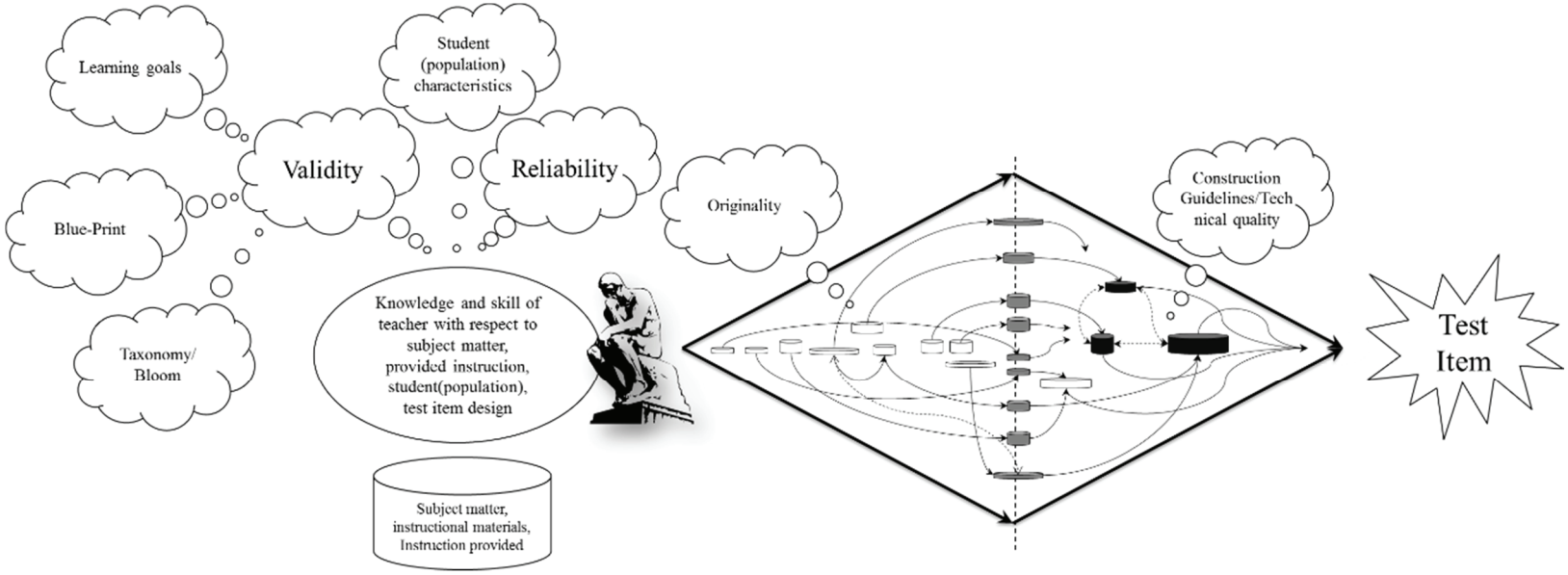

Figure 2. A visual representation of the concepts involved in test and test item design and the position of the teacher.

The thesis also contributes to research into computer environments for creative problem solving (Hewett, 2005; Mulet & Vidal, 2008; Shneiderman, 2002), as a concept for a computer support tool was developed. In this thesis, a clear rationale for the setup and elements of the tool were outlined, and could also be applied to other contexts. The concept could be of use in areas such as the field of web-based learning environments (Narciss, Proske, & Koerndle, 2007).

8.4 Strength of theory

An evaluation can be made regarding the extent to which the process model can be considered a *strong* theory. Prochaska et al. (2008) formulated a list of criteria that operationalizes the quality of theory in 12 fairly tangible criteria, stating that the more closely a theory fulfils those criteria, the stronger the theory is. The assumption by Prochaska et al. is that a strong theory can be well understood and communicated by being (1) clear, consistent, and parsimonious, but (2) also that the theory is generalizable by being integrative, testable, empirically adequate, and productive, and (3) the theory has important real-world implications by leading to utility, practicality, and discernible impact.

8.4.1 Communication: Clear, consistent, and parsimonious

First, an effort was made to develop a process model that was as clear, consistent, and parsimonious as possible, while drawing from multiple other theories and thus being integrative.

Clearness, consistency and parsimoniousness were maximized by developing relatively easily understood and visually representable concepts of diverging and converging as the main processes in test item design.

8.4.2 Generalizability: Integrative, testable, empirically adequate, and productive

The cognitive process model was firmly based on the knowledge base from several relevant domains concerning test item design, ill-defined problem solving, creative problem solving, and design problem solving. This knowledge was combined a new

model for a new type of problem. The model and the theory can therefore be regarded as a integrative.

The originating frameworks, models, and approaches on which the cognitive process model was based have already shown to be applicable and generalizable models in other domains and situations. Building on such existing findings offers a basis to support generalizability of the various findings.

Additionally, a firm distinction was made between diverging and converging, which allowed for specific empirical testing of the interventions with respect to these two processes. The theory was shown to be empirically adequate because the hypotheses of these studies were congruent with the evidence regarding improving the degree of originality of test items when the divergent phase of test item design was supported with specific interventions.
When the studies of Chapter 3 (Guidelines) and Chaper 4 (Design Patterns) of this thesis are considered case studies, they enable analytic generalizations (Yin, 2013) to other situations that occur in developing guidelines or design patterns. At a minimum, this can be concluded for the "hard applied" disciplines (Becher, 1994), because the case studies include a broad sampling of institutions, teachers, and specific domains and topics within the hard applied sciences. Additionally, test items for different purposes (exams, quizzes, activating learning materials) were developed. This is in line with the concept of variety sampling (Maxwell, 2012), in which as heterogeneous a sample as possible is sought for maximum coverage of different possible situations and contexts. With variety sampling, statistical generalization is not possible, but claims regarding analytic generalizations can be better supported. It is possible that the cognitive process model would be equally applicable within other domains or other contexts regarding test item design. With respect to broad disciplines as distinghuised by Becher (1994), the "hard pure," "soft pure," and "soft applied" disciplines (Becher, 1994), likely would be in need of different inspirational guidelines and design patterns.

Additionally, the theoretical model and findings from the experimental studies (Chapter 6 and 7) may be generalizable to test item design for teachers in primary and secondary education. The developed interventions were neither very domain -

nor context - specific. Additionally, the cognitive model and the interventions designed could yield the same value to test item design projects in which more resources are available, such as large-scale standardized testing projects.

Further, for the experimental studies, different groups of participants were used. This resulted in both a claim for the generalizability of the results but also in limitations. For the first experimental study (Chapter 6), first-year students were used. It may be assumed that this sample of students had some knowledge of the domain at hand, but they were obviously not as advanced as beginning teachers or experts (Chi, Glaser, & Farr, 1988; Ericsson, 2006; Regehr & Norman, 1996). This limits the generalizability of that study. However, for the second experimental study, participants were recruited from a population of junior teachers and teaching advisors who also had different prior knowledge of the topic for which test items were designed. That study therefore enabled a better generalization of the findings, especially because the level of expertise of the participants could be identified as a separate factor that had an influence on the outcome of the originality of the test items with respect to the divergent intervention. A possible limitation of the generalizability of the findings of the Chapter 6 and 7 studies is the relatively low number of participants, which limited the power of those studies and, possibly, leaving some mechanisms that influenced the test item design process undetected.

A limitation of the generalizability of this thesis is further that the hard pure sciences like math or statistics were not present in the studies. The hard pure sciences represent a domain which is more well-defined than in softer sciences or even the hard applied sciences. This could imply that the solutions space for generating test items would be more confined, resulting in a different perspective on the test item design task. The separate studies in this thesis did not discuss this difference.

Finally, the model could also be called productive because it revealed new phenomena and relations between the variables in the model. The model resulted in questions regarding the nature and importance of the divergent phase for test item design, support interventions to accompany the theory, and questions

regarding the relationship between originality and technical quality of test items, expertise, training, time-setting, exhaustion of ideas, extended effort, and computer environments to provide support. The model could is also productive in the sense that it can be used for other similar ill-defined problem types.

8.4.3 Implications for practice: Utility, practicality, and having an impact

This third group of criteria by Prochaska relates to the implications for practice of the developed theory. With respect to utility, practicality, and impact of the model, an effort was made to develop interventions, such as the inspirational guidelines, design patterns, a computer tool and creativity techniques, that supplement the theory. Measures were taken to support the claim that the results of the study therefore will lead to utility and impact.

First, studies concerning the development and evaluation of the inspirational guidelines and design patterns were conducted in the actual practice of teachers in higher education. These efforts elicited teacher's opinions on how to design test items in the domain of engineering and the life sciences and evaluate the possible utility of such guidelines for peers in their own community of teachers. The research thus strove to bring the findings directly to the intended target group, offering an immediate practical application and the promise of applications beyond the projects described. Further, the study regarding the design for a test item design support tool was undertaken in the form of a "design study";" rather than the "design-based research" or "educational design research" often used in educational research (Barab & Squire, 2004; Van den Akker, Gravemeijer, McKenney, & Nieveen, 2006), it adopted a research practice known from engineering and industrial design. Working with the design study approach explicitly enabled the study and definition of features and preliminary materializations of a tool which that could elicit first impressions regarding its expected usefulness and inspection of the adequacy of the rationale underlying the design.

Further, in the experimental studies, the tasks the participants had to perform were actual test item design tasks and not artificial cognitive tasks. Both the task setting and time restrictions to produce test items were set according to situations that

could be encountered in practice for higher education teachers. Additionally, as the developed interventions can be distributed to a wide audience with modern technology and with possible integration of those techniques in current computer-based testing systems, the impact could be substantial.

The criterion of impact is closely related to the question of whether the findings in this thesis will be used in the practice of higher education. Bringing about sustained educational change has proven to be a daunting challenge (Fullan, 2005; Henderson, Dancy, & Niewiadomska-Bugaj, 2012). The more radical a change is, the more effort is required to effectuate it; however, the degree of change is also directed to the chances for failure if the change does not comport with the received professional beliefs of teachers (Elton, 2003). Therefore, in this study it was an explicit aim to find innovative interventions in the form of support for teachers that improve the *existing* professional practices of designing test items for end-of-course in-house developed exams and activating learning materials (Elton, 2003; Miles, 1983; Schoonenboom, 2014). Further, when teachers are *adequately supported* in their existing professional practices, the chances are higher that teachers *sustain* any improved practice (Wieman, Deslauriers, & Gilley, 2013) and that improved final products (test items) will eventually be produced in larger quantities within the given context. In short, no far-reaching educational reforms were put forward in this thesis; rather, the emphasis was placed on addressing existing problems and on practical means to improve the performance of higher education teachers. The developed inspirational guidelines, the design patterns, and the creativity interventions were those practical means.

At the same time, presenting such improvements in a scaffolded manner, as by an educational technologist supporting a teacher, was observed to be important to raise the success in assisting teachers in the process of engaging with the support information. Letting teachers absorb the support information in accordance with their individual preferences and abilities or zone of proximal development (Vygotsky, 1978) weems crucial. Incorporating the findings of this thesis in item writing literature in teacher training programs and instructional designer training programs could help in this study making an impact in the future.

8.5 Further research

The studies in this thesis can serve as an initial step for further study and development of theory and practice. Next steps in research and development are needed and follow clearly from this thesis.

First, the validity of the inspirational guidelines and design patterns presented in Chapters 3 and 4 merit further study. For example, inspirational guidelines could differ for disciplines other than the "hard applied" subjects under study here. Actively engaging in finding and describing such guidelines could be of great value for those fields.

Second, in the experimental studies of Chapters 6 and 7, the tasks presented to the participants had very specific instructions on how to focus their efforts. Small differences in wording are likely to have a meaningful impact on outcomes for such tasks (Belson, 1981; Runco, Illies, & Eisenman, 2005). For example, if the tasks were formulated as "find the best single original test item," the results would differ from the instruction to "find the best test items to query for insight" or to "produce as many original test items as possible." Within the context of the experimental studies, the focus was on generating "as many good test items as possible." Therefore, the conditions created require careful attention by the readers of these studies to assess transferability to other circumstances. At the same time, this offers an interesting avenue for research; studies could be conducted to find differences in depth-first or breadth-first approaches to test item design (e.g., Tversky & Chou, 2011). In particular, studying whether a depth-first approach would lead to test items that had higher technical quality is promising, as the experiments showed a tradeoff between the number of test items developed and the technical quality of those items.

Third, it was found that participants of the experimental studies were only poorly able to optimize their original test items ideas to ensure clarity and unambiguousness of test items or adhere to the general rules. This is in line with findings from research regarding the difficulties of teachers in achieving that goal (Downing & Haladyna, 2006). The proposed presentation of the guidelines and the interactive texts in the online support tool did not result in a positive effect. An

important reason for that finding could be that it requires more consciousness-raising and skill among teachers to achieve it, which leads to questions on how to support teachers better in this task using online tools, in particular because that would lower the need for review of test items by peers or students (Haladyna, 2006). On the one hand, more elaborate experimental studies could be conducted to elicit detailed information regarding thought processes involved in optimizing test items for clarity and unambiguousness, such as by studying to what extent the cognitive processes described in Chapter 2 regarding competent teachers' inclination and ability to critique initial test item ideas from a variety of perspectives do in fact occur. On the other hand is the question of how this process of critiquing can be encouraged more effectively by online systems: what practical means, other than the general lists of rules can promote the inclination and effectiveness of critiquing items by teachers themselves? For that purpose, building further on the study of Mayenga (2009), who mapped teachers' ability to adhere to the item design task, is of interest. Perhaps incorporating "think aloud" protocol data would help determine how teachers deal with finding plausible distractors for test items or incorporating into the test item design tool user- or system initiated prompts to check test items more thoroughly while avoiding irritating the teachers.

A challenging further study would involve the utility of the developed online support tool and designed interventions to raise the originality of test items in actual practice *without* the supervision found in an experimental setting: how would the tools be used and would the study's results be found in the practice of higher education? The studies in Chapters 6 and 7 showed the utility of the support interventions in a controlled environment in which the participants were in a situation in which they could exert maximum performance. In practice, teachers control their own schedules and effort levels and they have to design a number of various test items for complete tests. Teachers could follow a route in which they design test items immediately before the test is administered or could design test items in a very methodical manner over a long period and be consciously gathering ideas throughout the instruction that leads up to exams. On the basis of such different contexts, different results could emerge. As the interventions cannot impose specific approaches by teachers, empirical evidence would be needed to

measure the use of the interventions and resultant changes in outcomes. The theory of the technology acceptance model (Davis, 1989) would be one means to study the possible use of a computer tool similar to Schoonenboom's (2014) study on how specific functions and features of technology-enhanced educational methods are used, based on perceptions of importance, usefulness, and ease of use by the teacher. Further study of the interventions based on this study is a highly promising line of inquiry.

8.6 In closing

Everyone who has attended college or university has without a doubt been subjected to repeated achievement tests, exams, and quizzes. Especially in large enrolment courses, the use of selected response test items is abundant and nearly every student has experience with some test items displaying flaws or a series of test items that query mainly for factual recall of information.

Though selected response test items are versatile and efficient artifacts to measure knowledge and skill, they are also the subject of substantial debate and criticism. Critique comes both from scholars (DiBattista & Kurzawa, 2011; Hansen & Dexter, 1997; Jozefowicz et al., 2002; Struyven, Dochy, & Janssens, 2005) and the general public (see for example Cizek, 2001; Cizek, Fitzgerald, & Rachor, 1995). The concern is often that test items contain indeed ambiguous information resulting in unfair tests, but even common is the concern that selected response test items primarily appear designed largely to call for recognition or recall of verbatim information without any context, leading to a superficial learning experience (Struyven et al., 2005). These criticisms are a pressing issue.

With respect to the second type of criticism, it has been reported that though teachers try to query for more engaging test items that query higher-order learning outcomes such as critical thinking or problem solving, they are not always able to design such items and thus often resort to test items that do indeed elicit only verbatim recall of factual information (Buckles & Siegfried, 2006).

From my own experience as a teacher at the Hague University of Applied Science, I can acknowledge this reality. Simply put, it is hard to design test items. I once

developed an item bank of six hundred true-false test items for a course on "Analysis of mechanical products" with a colleague. We worked very hard over many hours with many episodes of no progress; we were quite impressed with our own work when it was finished. However, having learned so much about test item design in the years since them, these items were in fact quite poor. The task of developing high-quality achievement tests, exams, and activating learning material consisting of test items requires skill, persistence, and inspiration. This problem is not alleviated simply by general item-writing training, as teachers have limited time and perform the test item design task without special assistance. How teachers can improve or sustain the quality of their practice and be supported to design more engaging test items remains a major problem, one that the advances outlined in this thesis can help correct.

From a theoretical and practical viewpoint, this thesis contributes to solving – or at least alleviating – this widespread problem in higher education. The developed process model acknowledges the intrinsically necessary cognitive processes involved in test item design from the perspective of the teachers and makes the nature and complexity of the task of developing test items more clear. The model can help in communicating purposefully with teachers about this process and provide more guidance in their task. In particular, the model provides leads and opportunities to improve divergent production of ideas for test items, an area much neglected in test item design. The work carried out in the studies in Chapters 3 to 7 have already resulted in ready-to-use inspirational guidelines, examples, support, and creativity interventions that can be used in the practice of every teacher in higher education. When the findings of the research reach individual teachers in higher education, it can contribute to the quality of testing and assessment in higher education.

8.7 References

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *Journal of the Learning Sciences*, *13*(1), 1–14.

Becher, T. (1994). The significance of disciplinary differences. *Studies in Higher Education*, *19*(2), 151–161. http://doi.org/10.1080/0307507941233138 2007

Belson, W. A. (1981). *The design and understanding of survey questions*. Gower.

Buckles, S., & Siegfried, J. J. (2006). Using multiple-choice questions to evaluate in-depth learning of economics. *The Journal of Economic Education*.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*(2), 121–152. http://doi.org/10.1207/s15516709cog0502_2

Chi, M. T. H., Glaser, R., & Farr, M. J. (Eds.). (1988). *The nature of expertise*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice*, *20*(4), 19–27. http://doi.org/10.1111/j.1745-3992.2001.tb00072.x

Cizek, G. J., Fitzgerald, S. M., & Rachor, R. A. (1995). Teachers' assessment practices: Preparation, isolation, and the kitchen sink. *Educational Assessment*, *3*(2), 159–179. http://doi.org/10.1207/s15326977ea0302_3

Cross, N. (2007). Forty years of design research. *Design Studies*, *28*(1), 1–4. http://doi.org/10.1016/j.destud.2006.11.004

Davidson, J. E., & Sternberg, R. J. (2003). *The psychology of problem solving*. Cambridge University Press.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–339. http://doi.org/10.2307/249008

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *The Canadian Journal for the Scholarship of Teaching and Learning*, *2*(2), 4. http://doi.org/10.5206/cjsotl-rcacea.2011.2.4

Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Elton, L. (2003). Dissemination of innovations in higher education: A change theory approach. *Tertiary Education & Management*, *9*(3), 199–214. http://doi.org/10.1080/13583883.2003.9967104

Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 683–703). Cambridge, UK: Cambridge University Press.

Fullan, M. (2005). The meaning of educational change: A quarter of a century of learning. In *The roots of educational change* (pp. 202–216). Springer.

Haladyna, T. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3–26). Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(3), 309–333. http://doi.org/10.1207/S15324818AME1503_5

Hansen, J. D., & Dexter, L. (1997). Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing testbanks. *Journal of Education for Business*, *73*(2), 94–97. http://doi.org/10.1080/08832329709601623

Henderson, C., Dancy, M., & Niewiadomska-Bugaj, M. (2012). Use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process? *Physical Review Special Topics - Physics Education Research*, *8*(2), 020104. http://doi.org/10.1103/PhysRevSTPER.8.020104

Hewett, T. T. (2005). Informing the design of computer-based environments to support creativity. *International Journal of Human-Computer Studies*, *63*(4-5), 383 – 409. http://doi.org/10.1016/j.ijhcs.2005.04.004

Isaksen, S. G., & Treffinger, D. J. (2004). Celebrating 50 years of reflective practice: Versions of creative problem solving. *The Journal of Creative Behavior*, *38*(2), 75–101.

Jacobi, P. (1997). *The magazine article: How to think it, plan it, write it*. Indiana University Press.

Jonassen, D. H. (1997). Instructional design models for well-structured and ill-structured problem-solving learning outcomes. *Educational Technology Research and Development*, *45*(1), 65–94. http://doi.org/10.1007/BF02299613

Jonassen, D. H., & Mandl, H. (Eds.). (1990). *Designing hypermedia for learning*. Springer-Verlag.

Jozefowicz, R. F., Koeppen, B. M., Case, S. M., Galbraith, R., Swanson, D., & Glew, R. H. (2002). The quality of in-house medical school examinations. *Academic Medicine*, *77*(2), 156–161.

Lipshitz, R., Klein, G., Orasanu, J., & Salas, E. (2001). Taking stock of naturalistic decision making. *Journal of Behavioral Decision Making*, *14*(5), 331–352. http://doi.org/10.1002/bdm.381

Maxwell, J. A. (2012). *Qualitative research design: An interactive approach*. SAGE Publications.

Mayenga, C. (2009). Mapping item writing tasks on the item writing ability scale. In *Canadian Society of Safety Engineering, XXXVIIth Annual Conference*. Carleton University, Ottawa, Canada. Retrieved from http://lib-ocs.lib.sfu.ca:8087/fedcan/index.php/csse2009/csse2009/paper/viewFile/1966/625

Miles, M. B. (1983). Unraveling the mystery of institutionalization. *Educational Leadership*, *41*(3), 14–19.

Mulet, E., & Vidal, R. (2008). Heuristic guidelines to support conceptual design. *Research in Engineering Design*, *19*(2), 101–112. http://doi.org/10.1007/s00163-008-0050-5

Narciss, S., Proske, A., & Koerndle, H. (2007). Promoting self-regulated learning in web-based learning environments. *Computers in Human Behavior*, *23*(3), 1126–1144. http://doi.org/10.1016/j.chb.2006.10.006

Prochaska, J. O., Wright, J. A., & Velicer, W. F. (2008). Evaluating theories of health behavior change: A hierarchy of criteria applied to the transtheoretical model. *Applied Psychology*, *57*(4), 561–588.

Regehr, G., & Norman, G. R. (1996). Issues in cognitive psychology: Implications for professional education. *Academic Medicine*, *71*(9), 988–1001.

Rowland, G. (1993). Designing and instructional design. *Educational Technology Research and Development*, *41*(1), 79–91. http://doi.org/10.1007/BF02297094

Runco, M. A., Illies, J. J., & Eisenman, R. (2005). Creativity, originality, and appropriateness: What do explicit instructions tell us about their relationships? *The Journal of Creative Behavior*, *39*(2), 137–148. http://doi.org/10.1002/j.2162-6057.2005.tb01255.x

Schoonenboom, J. (2014). Using an adapted, task-level technology acceptance model to explain why instructors in higher education intend to use some learning management system tools more than others. *Computers & Education*, *71*, 247–256. http://doi.org/10.1016/j.compedu.2013.09.016

Shneiderman, B. (2002). Creativity support tools. *Communications of the ACM*, *45*(10), 116–120. http://doi.org/10.1145/570907.570945

Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: a review. *Assessment & Evaluation in Higher Education*, *30*(4), 325–341. http://doi.org/10.1080/02602930500099102

Tversky, B., & Chou, J. Y. (2011). Creativity: Depth and breadth. In T. Taura & Y. Nagai (Eds.), *Design Creativity 2010* (pp. 209–214). Springer London.

Van den Akker, J., Gravemeijer, K., McKenney, S., & Nieveen, N. (2006). Introducing educational design research. In J. Van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational design research* (Vol. 1, pp. 3–7).

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Wieman, C., Deslauriers, L., & Gilley, B. (2013). Use of research-based instructional strategies: How to avoid faculty quitting. *Physical Review Special Topics - Physics Education Research*, *9*(2), 023102. http://doi.org/10.1103/PhysRevSTPER.9.023102

Yin, R. K. (2013). *Case study research: Design and methods*. SAGE Publications.